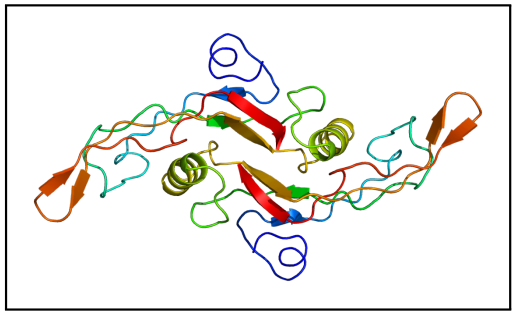
A clustering-based approach for predicting functional protein modules in TGF-β signaling pathways

Jean COQUET

Olivier DAMERON **Nathalie THERET**

Biological problem: TGF-β



- Transforming Growth Factor TGF-β:
  - Cytokine polypeptide
  - Three isoforms (1-3)
- Roles:
  - Regulate many biological processes: proliferation, differentiation, apoptosis...
  - Major role in the development of cancer

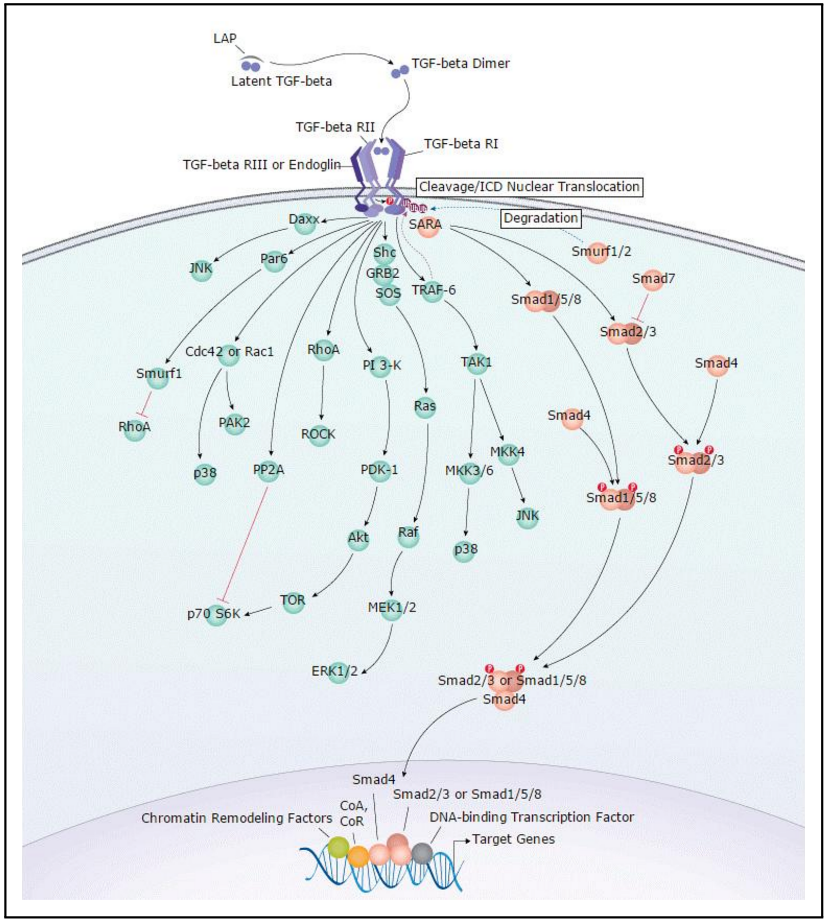
    Early stages: Anti-tumor role

    Advanced stages: Pro-oncogenic role

The pleiotropic effects of TGF-β are linked to the complex nature of its activation and signaling networks.

Biological problem: TGF-β

- The canonical (Smad) pathway
  1. TGF-β binds and activates receptors
  2. Receptors recrute and phosphorylate SMAD proteins
  3. Transcriptional regulation of target genes by SMAD complexes and transcription factors
- The non-canonical (non-Smad) pathways include JNK/P38 MAP kinase, Rho-like GTPase and PI3K.



Pathway Interaction Database

The Pathway Interaction Database (PID) was a collection of pathways composed of human molecular signaling and regulatory events and key cellular processes.

- 133 human pathway maps
- 9,248 reactions
- 27,876 biomolecules

Integration of PID database into a single unified model.

CADBIOM

CADBIOM = **C**omputer-**A**ided **D**esign for **BIO**logical **M**odels, developed by Geoffroy Andrieux.

http://cadbiom.genouest.org

Andrieux, G., Le Borgne, M., & Theret, N. (2014). An integrative modeling framework reveals plasticity of TGF-β signaling. BMC systems biology, 8(1), 1.

"CADBIOM is an open source modelling software. Based on Guarded transition semantic, it gives a formal framework for modelling cell signaling network."

Using CADBIOM language, the 133 signaling maps from PID were integrated in a unique discrete dynamic model (9,264 transitions, 9,177 molecules).

Analysis of TGF-β signaling networks using Cadbiom model

Find all chains of reactions linking TGF-β to gene transcriptions.

Results:

- **159** genes identified.
- **6,017 chains of reactions** linking TGF-β to at least one of the target genes in the nucleus.

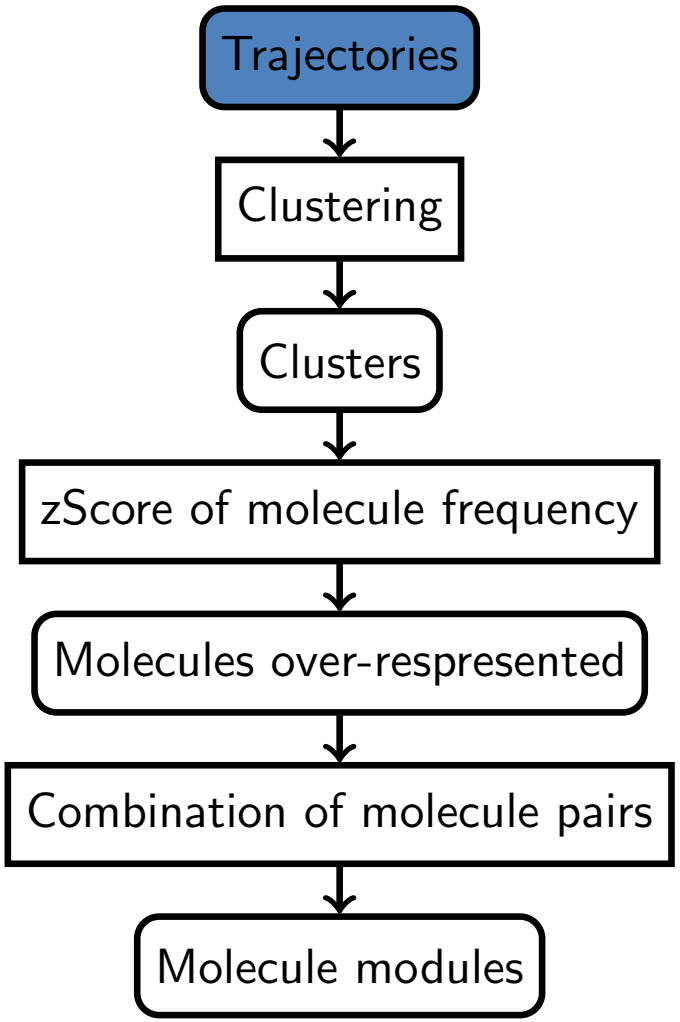
Definition:

- **Trajectory** = bio-molecule set "activated" initially or during the signal propagation and influencing the expression of a gene.

The complexity of trajectories

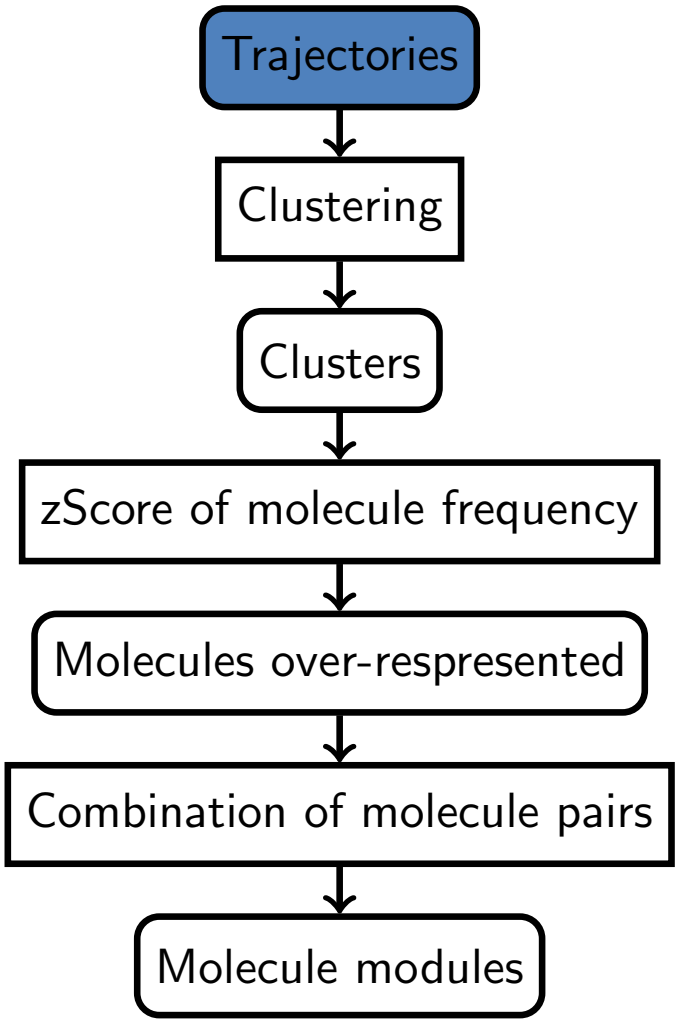
Figure: Representation of the 6,017 trajectories.

Characterization of the trajectories : A clustering based approach



- **6,017 trajectories** linking TGF-β to at least one of the target genes in the nucleus.
- **317 molecules** compose the trajectories.

Size and composition of the trajectories



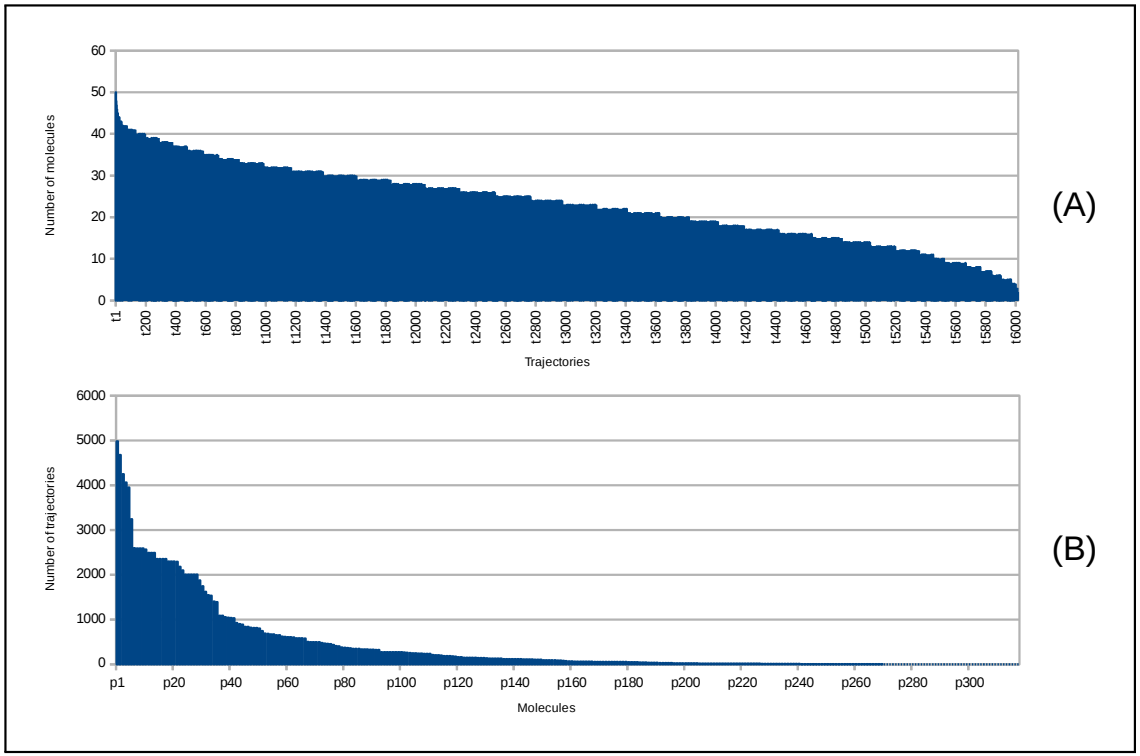
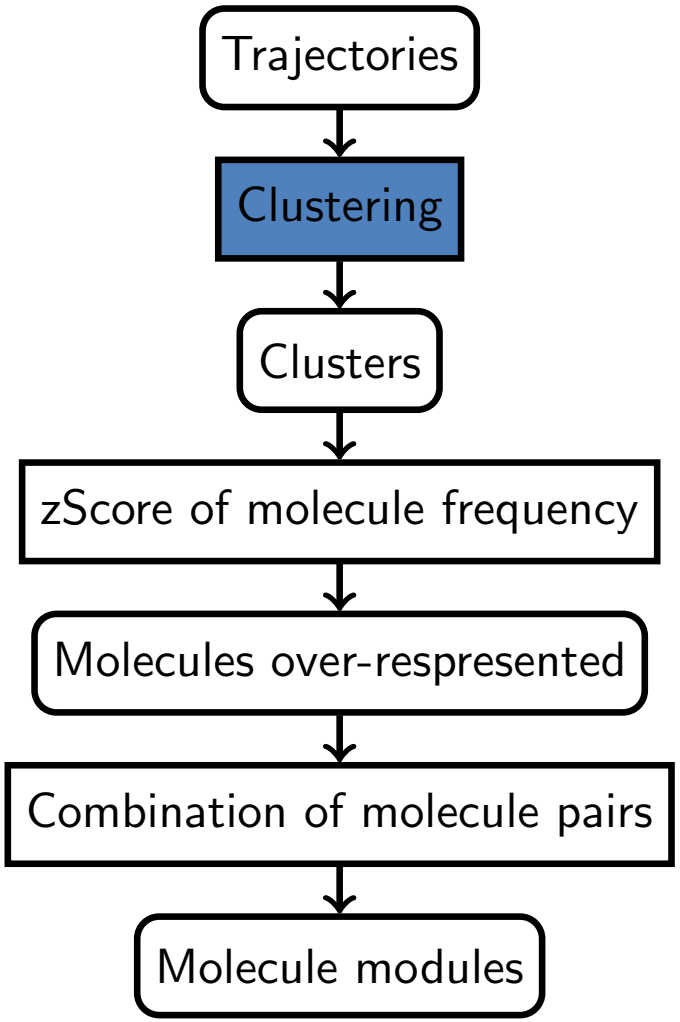


Figure: Distribution of (A) the number of molecules for each trajectory and (B) the number of trajectories involving each molecule.



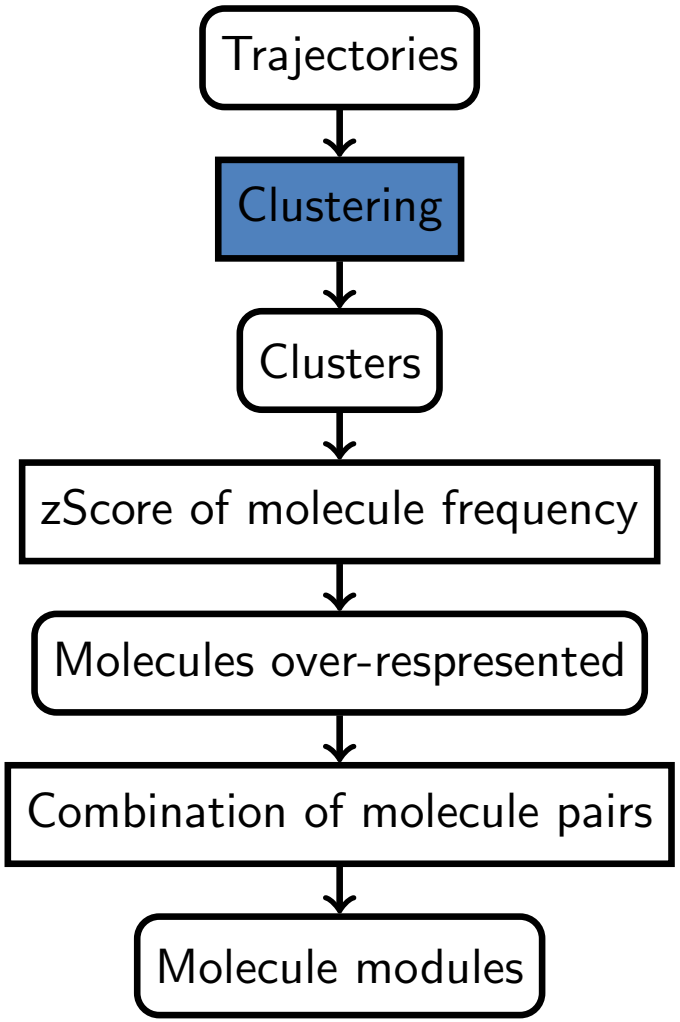
The Relevant-Set Correlation Model for Data Clustering.

Michael E. Houle.

2008.

Stat. Anal. Data Min. 1, 3 (November 2008), 157-176.

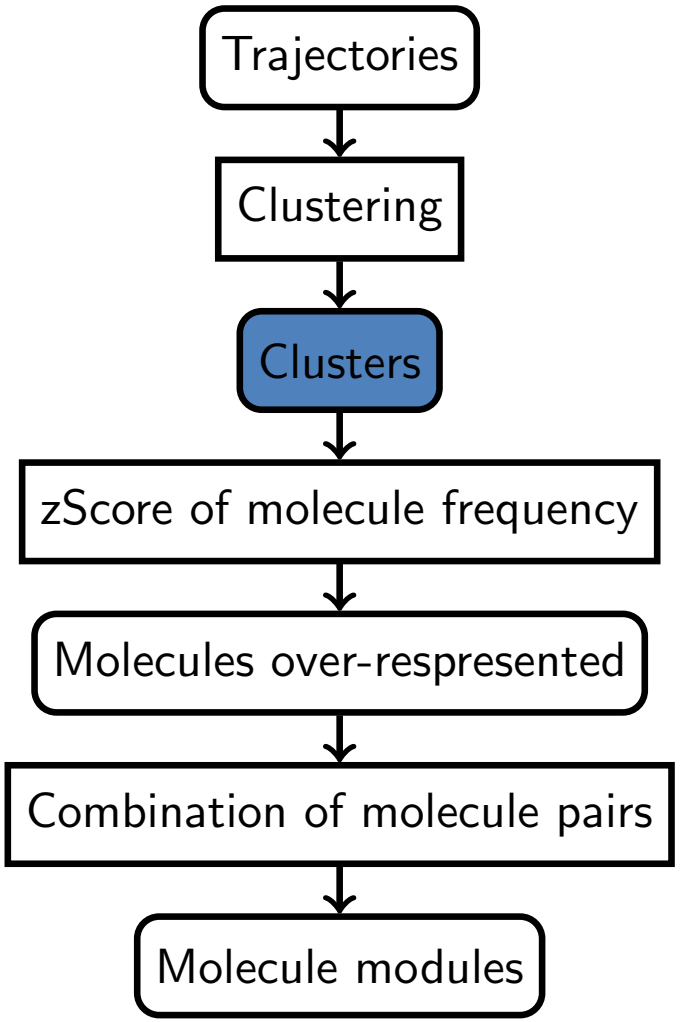
The Relevant-Set Correlation Model

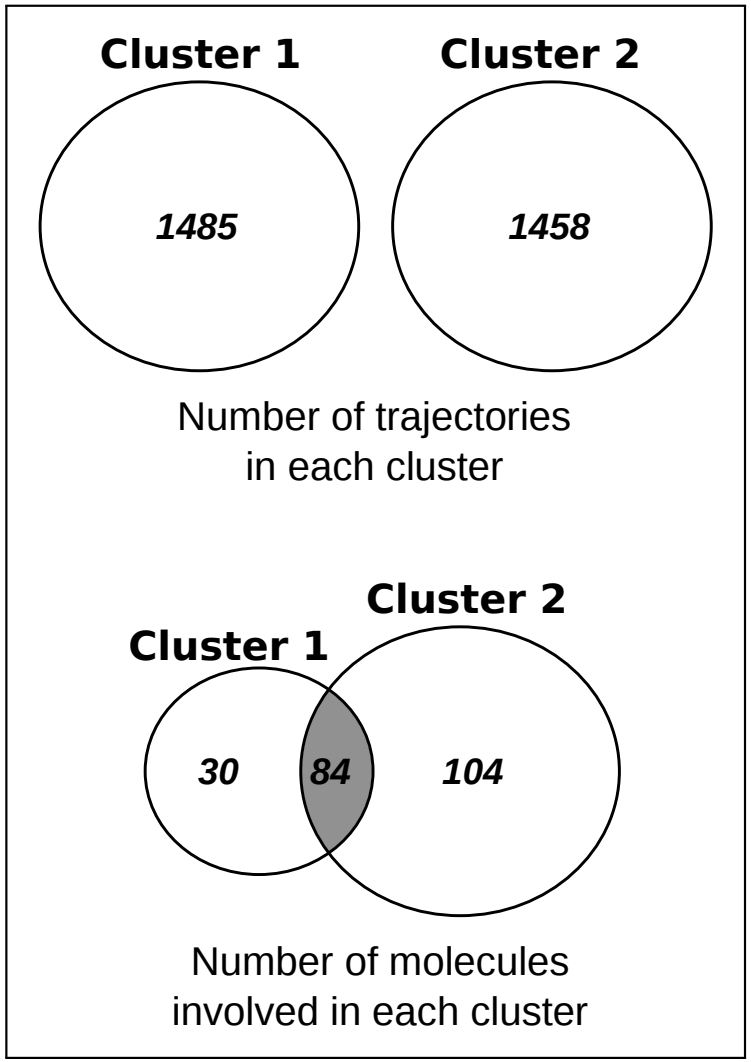


- Lets S the 6,017 trajectories and $t_k \in S$ one trajectory.
- With the Pearson correlation, we can compute the similarity between two trajectories based on the molecules they involve:

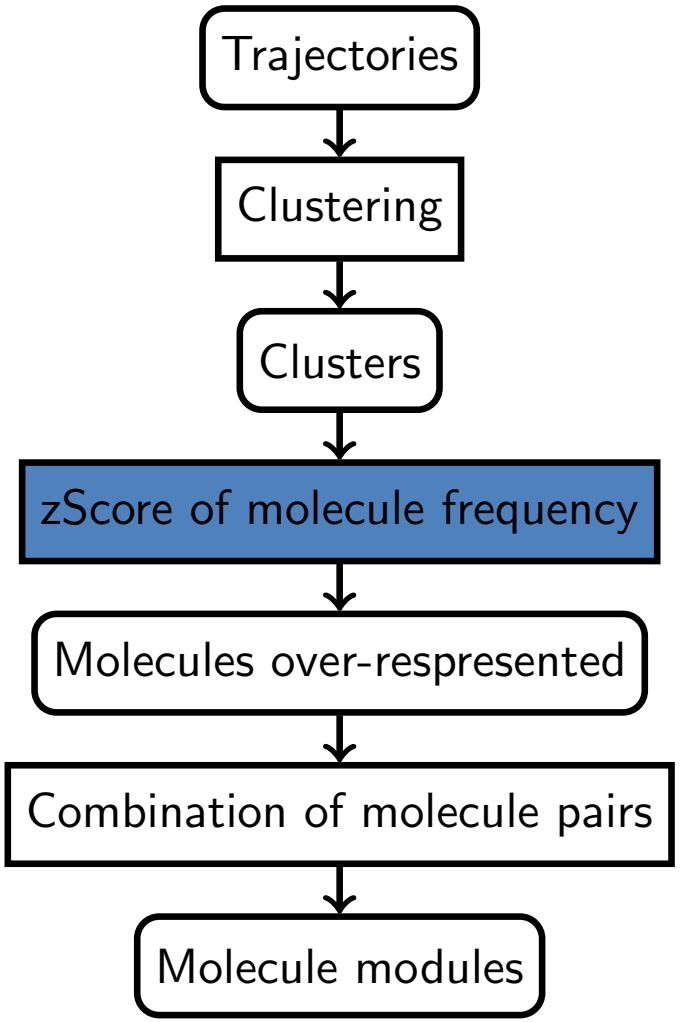
$$R(t_i, t_j) = \frac{|S|\frac{|t_i \cap t_j|}{\sqrt{|t_i||t_j|}} - \sqrt{|ts_i||t_j|}}{\sqrt{(|S| - |t_i|)(|S| - |t_j|)}}$$

Clustering results: two clusters





zScore of molecule frequency



Clusters are characterized by specific over-represented molecule signature.
Level of representation is computed by:

$$Z_A(p) = \frac{N_A(p) - F_S(p)|A|}{\sqrt{F_S(p)|A|(1 - F_S(p))}} \quad (1)$$

where :

- p is molecule
- A is a cluster
- $N_A(p)$ is the number of trajectories in A involving p
- $F_S(p)$ is the frequency of p in all trajectories S
- $|A|$ is the size of cluster

The clusters are characterized by opposite levels of molecules

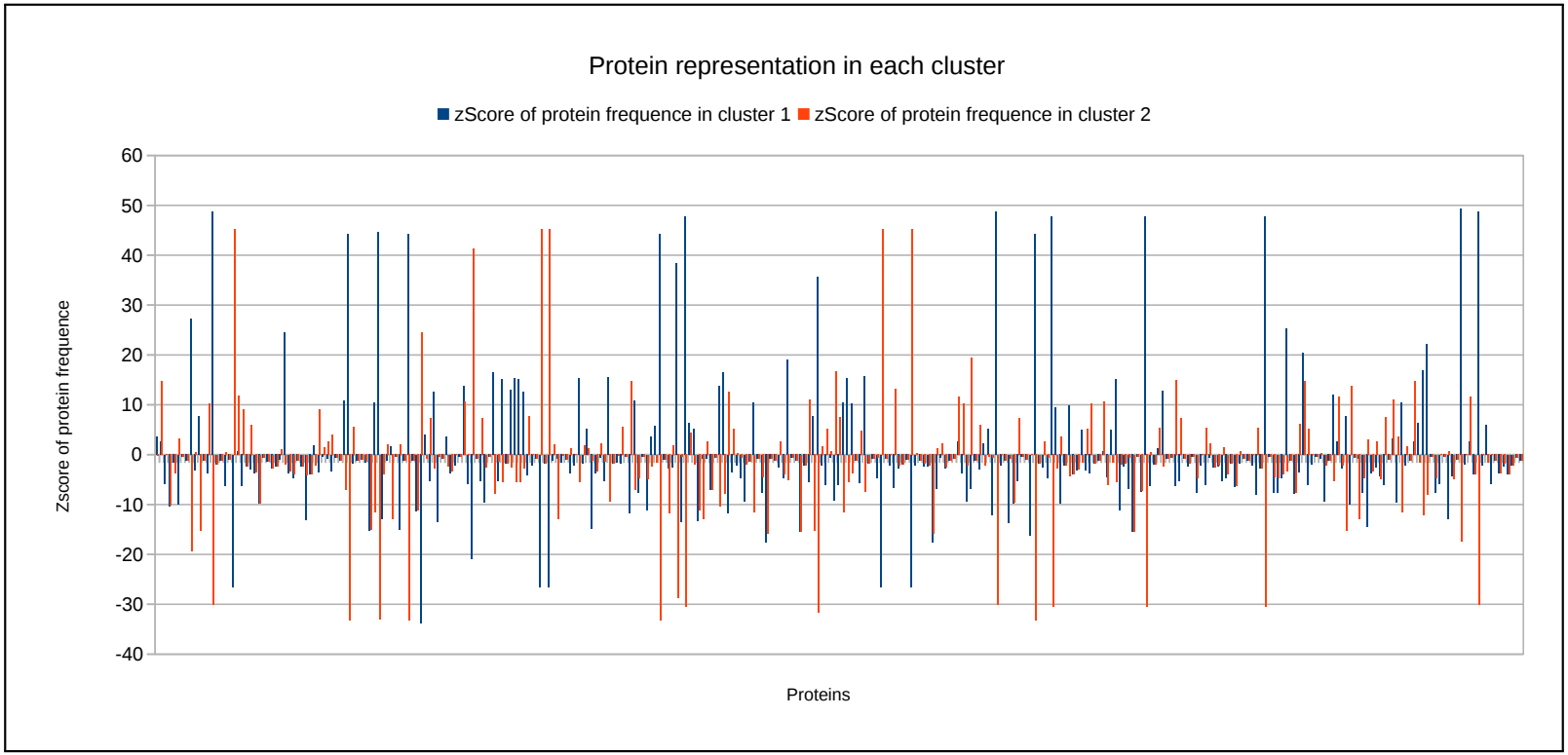
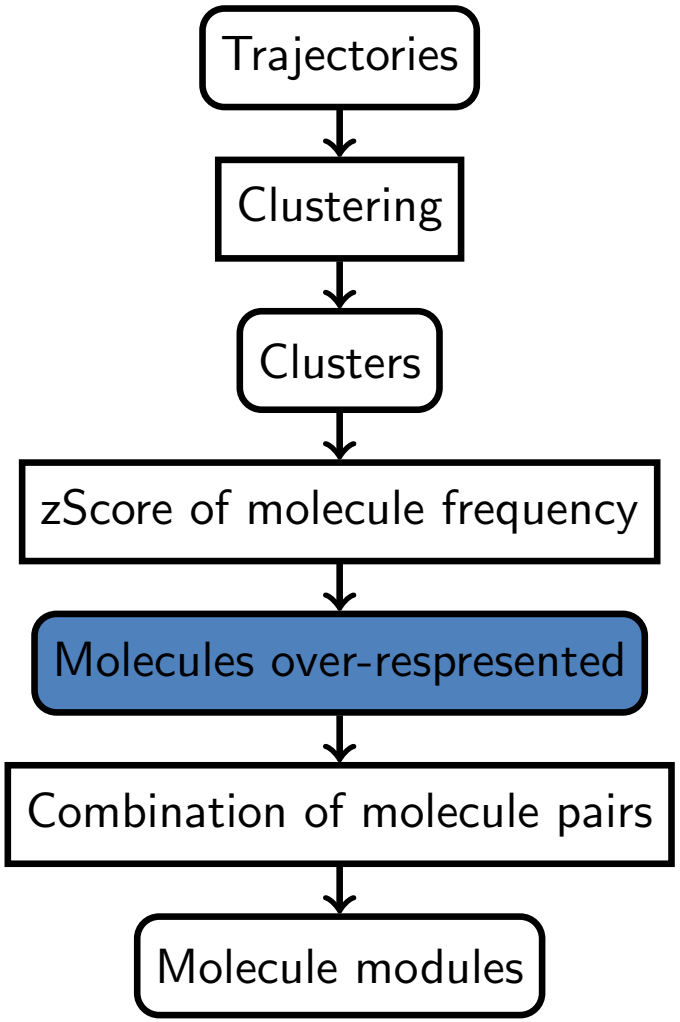


Figure: Representation of molecule levels

The clusters are characterized by opposite levels of molecules

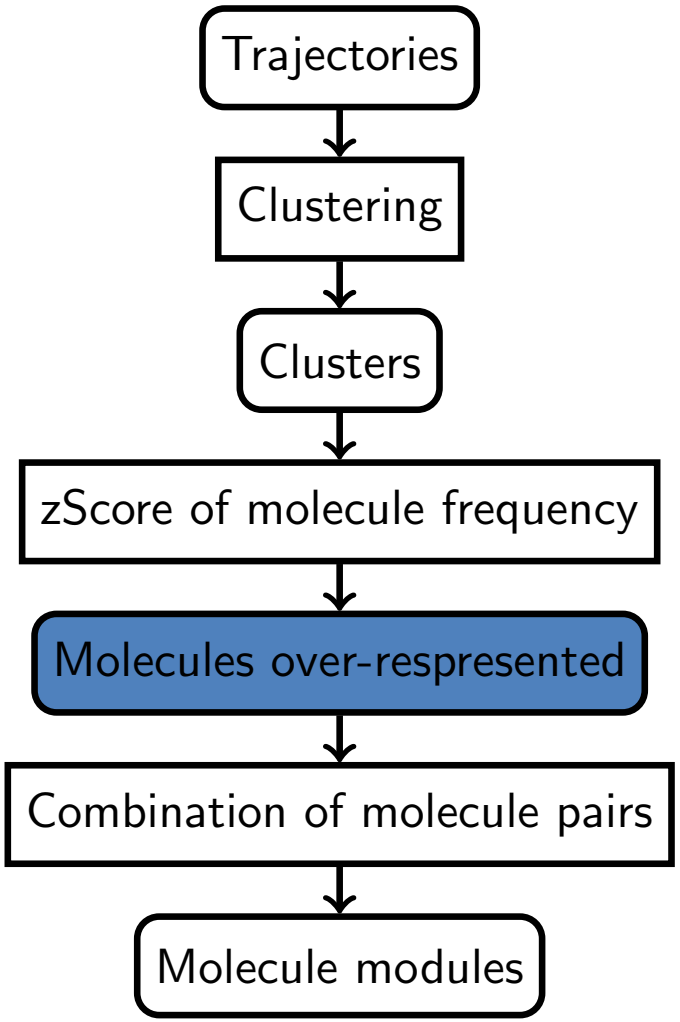


- Over-repesented molecules = molecules with a zScore ≥ 4.0
- Results:
  - **Cluster 1**: **61** over-repesented molecules
  - **Cluster 2**: **52** over-repesented molecules

⇓

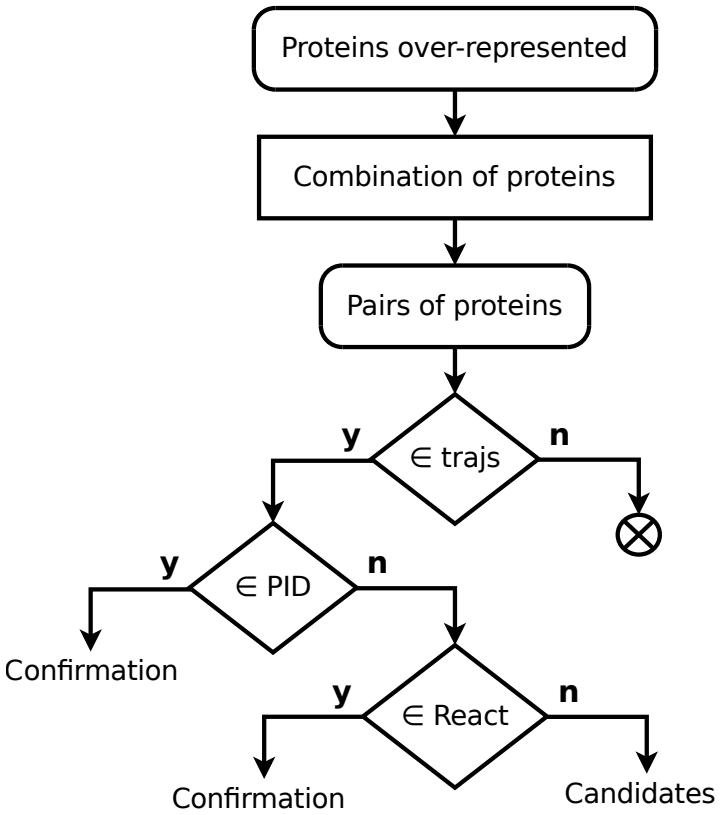
Is there a biological significance of the molecule signatures?

GSEA Analysis



- Gene Set Enrichment Analysis (GSEA)
- Results:
  - Cluster 1 = canonical responses of TGF-β including cell defense and immune innate response
  - Cluster 2 =
    - negative regulation of immune system
    - regulation of cell differentiation and development
    - epithelial cell proliferation
- Jekkyl and Hyde faces of TGF-β in cancer [Bierie and Moses, 2006].

Analysis of over-represented molecule pairs



	Core 1	Core 2
	61	**52**
	1830	**1326**
	1579 ∈ trajs	**980 ∈ trajs**
	1076 ∉ PID	**695 ∉ PID**
	529 ∉ Reactome	**398 ∉ Reactome**

Conclusion & Perspectives

Conclusion:

- We identified 2 families of TGF-b-dependent trajectories that summarize the antagonist effects of TGF-β.
- We extracted the signature molecules of each family.
- We found 529 and 398 protein pairs present in TGF-β-dependent trajectories that are never described in TGF-β signaling pathways from REACTOME database.

Perspectives:

- Implementation of new molecule modules influenced by the TGF-β in signaling database.
- Formulation of a new TGF-β signaling model.

Thanks for your attention!

Comparative analysis of over-represented molecule pairs with knowledge's from REACTOME database

Comparative analysis of over-represented molecule pairs with knowledge's from REACTOME database